Review of Selected Technical Characteristics of the Virginia Standards of Learning (SOL) Assessments

Ronald K. Hambleton, Chair
University of Massachusetts at Amherst

Linda Crocker
University of Florida

Keith Cruse
Texas Education Agency

Barbara Dodd
University of Texas at Austin

Barbara S. Plake
University of Nebraska at Lincoln

John Poggio
University of Kansas

The Virginia Standards of Learning (SOL) in the areas of English, mathematics, history, social science, and science are intended "to set reasonable targets and expectations for what teachers are expected to teach and students are expected to learn" (see the Virginia SOL Technical Manual, May 2000, page 1). The purposes of the educational assessments at selected grades (3, 5, and 8) and high school subjects are to inform parents and teachers about what students are learning in relation to the SOL and to hold schools accountable for teaching the SOL content. Therefore, this review has been undertaken only with these two purposes in mind; any other applications or uses of these assessments, such as for grade advancement or high school graduation, have not been considered.

The purposes of this report are (1) to summarize the views of the Virginia SOL Test Technical Advisory Committee (TAC) concerning several technical aspects of the educational assessments: Validity, Reliability, Equating, and Standard Setting, and (2) to identify several areas in which the TAC feels that additional technical work is needed. The TAC has met three times this year to discuss the technical aspects of the Virginia SOL educational assessments and has several more meetings scheduled for 2001. The Virginia Board of Education appointed the SOL Test Technical Advisory Committee in September of 1999. The TAC was charged with advising the Board of Education on technical matters related to the SOL assessments. Initially appointed by the President of the Board, subsequent vacancies and appointments will be made by the full Board.

Review of Selected Technical Characteristics

Validity

Validity is a property of scores obtained from an educational assessment, and concerns the extent to which the scores are useful for their intended purpose. Assessment of score validity is a critical element of any technical review. There is no limit to the amount of evidence that might be compiled to address the validity of the SOL scores in relation to their stated purposes and there is no amount of evidence that could ever prove that the scores from the assessments are valid. What is possible is the accumulation of a sufficient amount of evidence for the stated purposes of the assessments such that reasonable persons are willing to accept the validity of student scores and associated performance classifications (i.e., did not pass, proficient, advanced), until any counter evidence for the validity of the scores and associated decisions becomes available. It is also fair to say that all aspects of the assessment process, from preparing the SOL for guiding the test specifications, to writing test items, test scoring, test administration, test security, and standard-setting and equating, influence the validity of scores and associated performance category classifications from the assessments.

One important type of validity evidence for assessments such as the assessments in Virginia concerns the extent to which test items are providing information about the intended SOL. This is often called "content validity evidence" and there is ample evidence in the Technical Manual that the procedures used to investigate the content validity of the assessments were adequate. Fairly extensive procedures are in place to make these determinations using teachers from Virginia and measurement specialists.

Another aspect of content validity, sometimes called "domain validity," concerns the extent to which the current test specifications (these are the details about what the tests should measure and statistical criteria for the tests) are reflective of the SOL. That is, are the current test specifications broadly measuring the SOL? One possibility is that the current assessments are more than adequate measures of the test specifications, but that the test specifications themselves do not guarantee an adequate representation or sampling of the SOL. For example, some of the SOL calls for students to do an activity or interpret the meaning of a reading passage. Multiple-choice test questions may be well suited for some of the SOL, but may not be as effective in assessing other SOL. The SOL assessments are composed exclusively of multiple-choice questions. Therefore, it is possible that some of these standards are not fully reflected in these assessments. The TAC would like to see evidence of the extent of coverage of the SOL compiled in the future. The TAC would like to see documentation on the extent to which the test specifications match the SOL, and where they do not, we would be interested to know the steps that are in place for insuring that the areas of the SOL not covered on the assessments are taught and assessed in other ways.

Also, evidence is available in the Technical Manual suggesting that methods for identifying potentially biased items are in place. These are test items that are functioning differentially in subgroups of interest such as subgroups formed based on gender or ethnic background. Steps are taken to identify these problematic test items using both statistical and judgmental reviews. Standard statistical procedures are used, as well as committees of Virginia teachers. Both approaches for identifying these items functioning differentially in subgroups of interest are standard procedures in state assessments. More documentation in the Technical Manual, however, on the details of the statistical analyses would be helpful.

Documentation on the construct validity of the assessments is available in the form of correlations of SOL assessment scores with grades 4, 6, and 8 standardized achievement test scores (here, the Stanford 9 Achievement Test was used). Correlations between the SOL assessments and standardized achievement tests might be expected to be in the .50 to .80 region, and they are. The correlations are neither too high nor too low, lending some support for the validity of the SOL assessment scores. If the correlations were higher, then it might be said that the SOL assessments are measuring nearly the same knowledge domains and skills as the standardized achievement tests, and vice-versa. Validity evidence would be available to support the SOL assessments but it could be said too that the need for the Virginia assessments would be reduced. But the correlations should not be too low either because this may indicate problems with the reliability of scores on the SOL assessments, or suggest that Virginia is very much out of step in its curriculum frameworks with other states.

One assumption that is made when linking forms of the assessments across years and in score reporting is that each SOL assessment (e.g., Grade 3 mathematics) is mainly measuring a dominant trait or main factor. For example, in the mathematics area, the mathematics assessments ought to be measuring mathematics competence and not the combined effects of mathematics competence, reading/language proficiency, and the ability to work quickly. This assumption is often called the requirement for "test unidimensionality." Evidence for the validity of the unifactor or unidimensionality assumption with an assessment can come from reviewing the "eigenvalue plot" that is obtained from analysis of the correlation matrix formed from the correlations of all pairs of items in each assessment. The evidence for test unidimensionality among the Virginia SOL assessments is strong.

Reliability

The TAC reviewed the reliability evidence for the 1998, 1999, and 2000 administrations of the Virginia SOL assessments. Reliability is a characteristic of the scores and concerns, in the case of the Virginia assessments, the internal consistency of scores and the consistency and accuracy of assigning students to performance categories. For example, it is important to demonstrate that student scores are consistent over short-time periods and content samples, and that the performance category classifications being made with the scores are consistent and accurate over (say) two administrations. A well-known and widely accepted reliability statistic ("Kuder-Richardson Formula 20 (KR-20)") was used to estimate the internal consistency reliability of each grade and form of the multiple-choice SOL assessments. A second statistic (a less well-known statistic, "person separability index") was used to estimate the reliability of each grade and form of the writing assessments. Standard errors of measurement were also reported for each grade and form of the assessments.

The KR-20 coefficients have changed little across the three testing years. (Note that the KR-20 statistic has a range of values between .00 and 1.00, and values above about .80 are desirable with achievement tests.) The majority of the KR-20 coefficients for the SOL forms range from .85 to .92. The lowest coefficient across the assessment years was .81 for one of the two forms in science at grade 5. The highest reliability coefficients (.92) were obtained in mathematics at grade 8. In general, the high school end-of-course assessments had slightly higher KR-20 coefficients (.87 to .91 for spring 2000) than the SOL forms for earlier grades. Overall, the KR-20 coefficients revealed reasonably high reliability of the SOL multiple-choice forms. The person separability indices for the writing assessment ranged from .82 to .88 in 1998, .83 to .89 in 1999, and .84 to .88 in 2000. As was the case with the multiple-choice SOL assessments, the highest reliability coefficients were obtained for the high school writing assessment (.86 to .89). The TAC recommends that coefficient alpha be used to assess reliability of the writing assessments in the future. This change to coefficient alpha will allow the reliability coefficients obtained for the writing assessments to be compared to the KR-20 coefficients used to assess the reliability of the multiple-choice SOL because KR-20 is simply a special case of coefficient alpha.

In reviewing the descriptive statistics that are reported in the tables that contain the reliability coefficients in the Technical Manual, the TAC suggests that the mean raw score should also be reported as a proportion of maximum scores. This change would facilitate the comparison of assessments that vary in length for the different subject areas. Also the conditional standard error of measurement should be reported at the Proficient and Advanced performance standards.

It is the sizes of the errors in scores near the performance standards that are of special interest when interpreting scores and evaluating the suitability of the assessments for achieving the desired purposes.

Decision accuracy and consistency indices were also reviewed for the 1998, 1999, and 2000 assessments (these statistics, too, range from .00 to 1.00). We were especially pleased to see these statistics being reported because they are directly related to the stability of student performance category classifications resulting from the assessment scores. The accuracy of the decisions made between Passing (Proficient or Advanced classifications) and Not Passing for all grades across the three administrations ranged from .87 to .93. These coefficients are sufficiently high to justify the use of the assessment scores in performance category classifications. The index of the consistency of the decisions made between Passing and Not Passing at each grade and form were also acceptable (.80 to .89 in 1998, .82 to .91 in 1999, and .82 to .90 in 2000).

Many of the statistics mentioned above are provided for each form of the assessment for each grade level. The forms are labeled Core 1 and Core 2 where Core 1 is the primary assessment taken by the vast majority of the students and Core 2 is the make-up assessment taken by students who missed the administration of the Core 1 assessment. The TAC suggests that a description of the test development process that produced these forms should be included in the Technical Manual.

Evidence is also provided in the Technical Manual concerning the inter-rater reliability of the scoring of the writing prompts. These reliability estimates are strong suggesting that training and quality control on the scoring are adequate.

Taken collectively, the reliability evidence for the SOL assessments is solid and is typical of high quality assessments. In general, the assessments meet or exceed the reliability standards for such assessments. The recommendations of the TAC for changing the statistics used to assess the reliability of the writing assessment and some of the content of the Technical Manual are meant to increase the usefulness of the information to users interested in the technical aspects of the assessment. We are pleased to note that a number of our suggestions from an earlier meeting have already been incorporated into the revised Technical Manual.

Equating

Equating or test score equating is a statistical procedure that must be carried out to insure that the Core 1 and Core 2 assessments in a single year, and from one year to the next, are statistically equivalent. Assessments of equivalent difficulty are needed to insure fairness to students, and to insure meaningful interpretations of the assessment results from one year to the next. Certainly the test developer tries to construct assessments of equivalent difficult but this turns out to be rather hard to do because of a finite number of items available to build forms and the necessity of using pretest item statistics. Equating of scores is common in the testing field. There is probably not a state assessment in the country that does not do statistical equating of scores across multiple forms, and from one year's assessments to the next.

The TAC is concerned that this rather complicated activity of equating must be carried out at the end of the first two weeks of testing each year so that scores can be reported to those taking the assessments in the first two weeks of the assessment window in a timely fashion. Besides the complexity of the analyses, and the speed with which the work must be done, there is the danger that running these equating analyses on students taking the assessments in the first two weeks may produce results that might not be achieved were the full sample of students used or a statistically representative sample of students. The TAC strongly recommends that more research be carried out on the suitability of the sample of students taking the assessments in the first two weeks for equating the forms. We were pleased to see at our last meeting that this important line of research is well under way by the Department and its contractor.

Perhaps a few more specifics on this potential equating problem would be useful. The equating procedure used to equate Core 1 and Core 2 forms in a given assessment year and the equating of assessments across years are done with a "post-equating design." Post equating in Virginia is hampered by the administrative constraints of the program. The five- week staggered administration of the assessment and the 14-day turn-around on scoring means that the equating sample might not be representative of the entire student population. Only students who are assessed early are included in the post-equating process. The committee views this as a potential weakness in the program. Given the high-stakes nature of the testing program, the committee suggests that other equating designs should be investigated. Specifically, the committee recommends a pre-equating design be investigated to alleviate some of the problems associated with using only data from schools that administer the assessments early in the five-week testing window. Also, while rapid turn around of results may be desirable, the condition that score reports be returned to schools in 14 days is a scenario that is fraught with potential quality control issues. We would like to see the requirement for two-week turnaround of scores reconsidered.

Standard Setting

Standard setting is the process of determining the levels of achievement required of students to be placed into the performance categories of interest. In Virginia's SOL, there are three performance categories, labeled, "Did not pass," "Proficient," and "Advanced." The TAC's view is that acceptable methods were used to set the performance standards. There are sufficient details available also about the process for full reviews to be carried out. The state department and its contractor have chosen and implemented two methods for standard setting (the Angoff method and the Bookmark Method) that are acceptable for multiple-choice assessments and the evidence seems to be that the methods have been appropriately implemented. At the same time, the TAC also recommends that research on the validity of the performance standards be continued because of their central role in the assessment program. Here are a few sample questions: Is there evidence that the same performance standards would be obtained if different panelists were used? Is there evidence that teachers would assign students to the same performance categories as they are assigned using the assessments and performance standards? To what extent are students who score near the performance standards differentiated on other suitable criteria from students who score some distance from the performance standards?

Follow-Up Technical Work

In our review, the TAC has identified a number of topics that require further study. First, now that the SOL and associated educational assessments have been in place for three years, it is timely to begin to investigate the consequential validity of the assessments. A complete agenda for consequential validation should be developed carefully with input from stakeholders and educators. The following questions are suggested as illustrative, rather than as prescriptive, for the type of evidence that should be gathered. What evidence is there that the Virginia SOL is having an impact on education? Here are a few illustrative questions that the Department might want to consider in the coming years: What percent of teachers are now fully committed to teaching the curricula? Is there any evidence that the curricula are being narrowed to focus only on the parts that are included in the assessments? Are parents any more involved in education, than (say) four or five years ago prior to the program and are the opinions of parents and other stakeholders addressed? Are the state teacher-training programs doing a better job of preparing teachers? How have the teacher training programs in the state changed their approaches, if at all, for preparing teachers? Are more students going on to college now? Are the colleges noticing that the students from Virginia schools are more qualified than before implementation of SOL assessments? Is there any evidence that the five-week window for test administration is giving an unfair advantage to schools testing later because of security breaches? What are the identifiable features and educational practices of schools that score high on the SOL? These and related questions need to be studied to determine both the positive as well as negative consequences of the SOL, including both anticipated as well as unanticipated outcomes. The TAC notes that it would have been premature to carry out these studies earlier because time is needed for the SOL to take hold. But the time has arrived for a program of research to investigate consequential validity. What the TAC recommends at this time is that the Department prepares a research agenda spanning the next several years.

Second, the Technical Manual needs to be expanded to include more descriptions of procedures, results, and statistics. The current draft is well done, but more can always be done, and we recommend that expanding the Technical Manual become a priority in the coming year. There should be sufficient details that other researchers could replicate the procedures used in Virginia. Such details allow for a full technical review. In particular, more details are needed on the content review of items, on the bias reviews, on the standard errors around the performance standards and on the details of the standard-setting process. For example, with the standard-setting, validity evidence such as panelist impressions of the methodology and variability of performance standards across panelists would be helpful. The TAC was generally pleased with the May 2000 version and more pleased with the later version that was released recently. A number of our requests had been addressed in the updated Technical Manual. The Department's responsiveness was very much appreciated by the TAC.

Finally, we are very concerned about the current design for equating of Core 1 and Core 2 forms each year, and forms from one year to the next, because of the need to equate forms using the early test-takers each year. Flaws in the equating will have a direct negative impact on the validity of scores from the SOL assessments. There is the danger that the earlier test takers each year are not fully representative of students across the state, and so this unique sample of test-takers may produce a linking of forms that would not hold up were it carried out with more representative samples. New designs need to be considered. The problem is created by the five-week window of available test dates, and the need to turn test results back to schools within two weeks of test administration. These conditions impose constraints on linking designs and test security that are highly problematic and so the conditions themselves need to be studied too. At our meeting on October 13, 2000, we were pleased to see that important research has been initiated on this topic, and already has produced some useful results.

Summary

Based on our review to date, the TAC has concluded that the Virginia Department of Education and its contractor are following standard procedures for the design and implementation of state testing programs. The TAC was generally pleased with the technical quality of the educational assessments. There is evidence that the educational assessments are being developed in a professionally responsible way; there is evidence of content and construct validity; both score and decision consistency are high; and performance standards were set in a defensible way. At the same time, research to investigate consequential validity is very much in order in the coming year, better technical documentation is needed, and the problems of linking forms from one year to the next need to be further investigated.